Intrinsic Motivations for Forming Actions and Producing Goal Directed Behaviour

Daniele, Caligiore; Das, Gautham; Valerio, Sperati; Varun, Kompella; Vieri, Santucci; Benureau, Fabien; Francesco, Mannella; Mai, Nguyen; Marco, Mirolli; Vincenzo, Fiore; Gianluca, Baldassarre; Robert, Nawrocki

Publication date:
2011

Citation for published version (APA):
Daniele, C., Das, G., Valerio, S., Varun, K., Vieri, S., Benureau, F., Francesco, M., Mai, N., Marco, M., Vincenzo, F., Gianluca, B., & Robert, N. (2011). *Intrinsic Motivations for Forming Actions and Producing Goal Directed Behaviour*. Paper presented at Capo Caccia Cognitive Neuromorphic Engineering Workshop, Aberystwyth, United Kingdom of Great Britain and Northern Ireland. http://hdl.handle.net/2160/7586

Intrinsic Motivations for Forming Actions and Producing Goal Directed Behaviour

Fabien Benureau, Gautham P. Das, Varun Kompella, Robert A. Nawrocki,
Sao Mai Nguyen, Gianluca Baldassarre, Marco Mirolli, Valerio Sperati,
Francesco Mannella, Vincenzo Fiore, Daniele Caligiore, Vieri Santucci

May 20, 2011

Abstract

In classical reinforcement learning framework, an external, handcrafted reward typically drives the learning process. Intrinsically motivated systems, on the other hand, can guide their learning process autonomously by computing the interest they have in each task they can engage in. We explore how intrinsic motivation could be implemented in the iCub platform on a learning task that was used previously with infants and monkeys, with a focus on discriminating between task of varying difficulty, and observing how their interest towards the tasks change as their knowledge of them progresses. Two main different approaches were taken : one where the reinforcement learning framework was adapted to an intrinsic reward, and another where the focus was put on a goal-oriented architecture. Two experiments settings were used, one with a console proposing buttons that activated boxes, and another proposing an interaction with rods : both experiments exhibited two tasks, one easy, and one difficult to learn. In each experiment, the system is able to successfully focus on learning the easier task earlier than the difficult one.

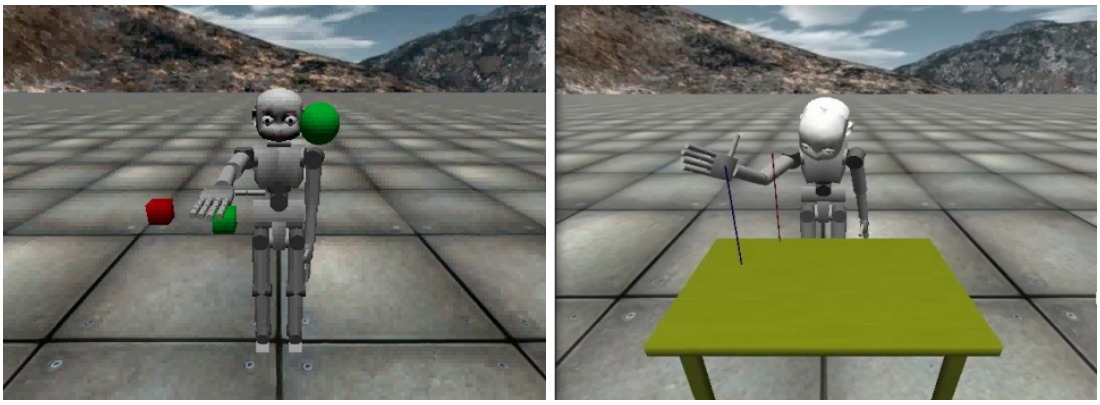
1 Introduction

Most of the initial researches in robotics were focused on controlling robots in structured environments, handling a set fixed objects and tasks. Tomorrow humanoid robotic systems are expected to seamlessly integrate in human environments, carry tasks and assist humans. The challenges that such a perspective represents are many, and one of the requirements of such robotic systems is to be able to learn new skills autonomously, such as walking on unknown surfaces, manipulating new objects or handling unexpected requests or situations.

With a robot of the complexity of the iCub, fitted with numerous, high-resolution sensors and actuators, the set of tasks and skills to learn is so large that exhaustive learning is impossible, even at the timescale of a lifetime. Moreover, some tasks are more difficult to learn, and some skills will be of considerable use in the typical activity of the robot (walking, grasping), while other will be of little to no use (touching its nose).

Providing the agent with an external reward signal has proven to be a very effective learning method in a variety of situations. Yet, such a reward typically has to be handcrafted at design time, and is hardly compatible with the unpredictable nature of the tasks the robot may encounter. A solution to this problem is intrinsic motivation : we provide the robot with the capability of expressing varying amount of interest in the different tasks it faces, mimicking the human and animal curiosity. If the interest measure of the robot is effective, then useful and learnable tasks will get a high level of interest. The interest measure is then used to select which task to focus on.

Intrinsic motivation has drawn a lot of attention recently, especially for open-ended cumulative learning of skills Weng et al. [2001], Lopes and Oudeyer [2010]. The word Intrinsic motivation was first used in psychology to describe the capability of humans to be attracted towards different activities for the pleasure that they experience intrinsically. These mechanisms have been shown crucial for humans to autonomously learn and discover new capabilities Ryan and Deci [2000], Deci and Ryan [1985], Oudeyer and Kaplan [2008]. This inspired machine learning researchers and developmental roboticists in creating



(a) The iCub Simulator with the experimental setting. Here the iCub is pushing the green button (cube), opening up the green box (sphere).

(b) Experimental setup for the rod toppling experiment. Robot interacts with the rods by toppling them off on the table.

Figure 1: iCub Experimental Scenarios

fully autonomous robots Barto et al. [2004], Oudeyer et al. [2007], Baranes and Oudeyer [2009], Schmidhuber [2010], Schembri et al. [2007]. Since the onset of their investigation, they have proposed numerous ways of integrating such systems on the basis of meta-exploration mechanisms monitoring the evolution of learning performances of the robot, in order to direct it towards regions of the space where it would obtain maximum informational gain. Heuristics defining the notion of interest were studied in active learning Fedorov [1972], Cohn et al. [1996], Roy and McCallum [2001].

If the interest measure of the robot is effective, then useful and learnable tasks will get a high level of interest. The interest measure is then used to select which task to focus on. In this report, we focus on designing interest measure and exploring how such an interest measure can discriminate between tasks of varying difficulty.

2 Experiments

Two different scenarios with multiple tasks of varying complexities are considered for the evaluation of the intrinsic learning models. In both the scenarios the actions are associated with the movement of the hand of iCub robot to do different tasks.

2.1 Button Experiment

Figure 1a shows a simulation of the experimental setup, where the iCub is placed in front of a control board with five buttons. Each button can take different colour through coloured lights. At any moment during the experiment, two buttons are activated, one in red and another in green; the others are off. The robot has five possible actions, each corresponding to a button at a fixed location. The action includes the arm movement from the initial position to the button, the button press and retracting the arm back to the original position. If a button with a specific colour is pressed, a box of corresponding colour opens. The experiment is sequenced in trials; during each trial the iCub can experiment with the result of each button push, until a timeout is reached.

From one trial to another, the position of the red button doesn't change. However, the position of the green button is selected randomly among the four remaining position. Therefore, while managing to open the red box is quite easy to learn, opening the green one reliably between trials necessitates extended learning. The goal is to develop an intrinsically motivated learning algorithm that focus first on learning the easy task and then one the more difficult one. If such algorithm can be generated for complex general goal spaces, it is hoped that selectively focusing on tasks of increasing difficulty will, first, provides the robot with skills at any point during the learning (it's better to know well few things than almost nothing about everything), and also potentially allow the emergence of an autonomous scaffolding behaviour in the learning process.

2.2 Rod Toppling Experiment

We discuss here another experimental setting where the iCub robot learns to interact with two rectangular objects (blue and red rods) on a table. Figure 1b shows a snapshot of the simulation environment. The robot can move its right hand to 5x3 grid positions (termed as states) over the table with four possible actions: forward, backward, left and right. The red rod is made more easily accessible than the blue rod, i.e., there are more state-action pairs to topple the red rod than the blue rod. The robot tries to predict the state of the rods (whether toppled or not) for each of the visited state-action pairs of its right hand and in the process learns to build action-sequence to topple the rods. A tabular Q reinforcement learning algorithm (see Section 3.3) is used for control and a tabular linear predictor is used to predict whether the object is standing still or is-toppled.

3 Intrinsically Motivated Reinforcement Learning

Extrinsically motivated reinforcement learning involves learning through trial and error so as to maximise reward from the environment. The extrinsic reward is obtained when the task is successfully accomplished. The application of extrinsically motivated learning in a robot has limited scope as the reward has to be defined for each task, while the system is designed. This limits the learning capability of the system to limited set of tasks.

Intrinsically motivated reinforcement learning doesn't involve any reward, where as the changes in the environment resulting from the actions form the curiosity which will drive the learning. Though the system is difficult to model, it can be used for learning more tasks as the reward for each task need not be defined a priori. Implementing an intrinsically motivated learning system in a humanoid robot improves its adaptability in unstructured environments. Intrinsic motivation can also be implemented within the framework of Reinforcement Learning Sutton and Barto [1998], which aims in general at the learning by the agent of a particular task in interaction with his environment. Among Reinforcement Learning algorithms, Temporal-Difference approaches enable a non-model-based learning of the environment. We explore how two particular algorithms of the TD approach, the Actor-Critic algorithm and the Q-learning algorithm, can integrate intrinsic motivation.

3.1 Goal-Oriented Architecture

In order to focus on the creation of efficient interest measure, we created a simple modular architecture for our agent (figure 2) : a predictor, an actor, and an interest module. Figure 2 shows the different modules in this approach and the interactions. The numbers indicate the order of the module interactions. First (1), the actor gets an observation of the environment. Then (2), the actor asks the predictor for the effect an action would have in the current context. This effect is transmitted (3) to the interest module, which expresses to the actor (4) how interesting achieving this effect (hence, the corresponding action) currently is. The cycle (2,3 and 4) repeats for any action the actor considers. The actor then chooses an action taking into account its interest, and executes it. The environment state is affected, and the global cycle repeats.

Prediction Module

The prediction module stores past experiences, and given a context and an action provide prediction of the effect produced. It provides an estimation of the transition function $S \times A \rightarrow S$, where S is the space of context, and A the one of actions. For the modest purposes of our experiment, our prediction module is a simple nearest neighbour research on past experiences, assuming a non-stochastic, markovian property of the environment.

Actor Module

The actor module request the interest value for each of the five actions given the current context, and then use a ϵ-greedy selection method to chose which one to execute.

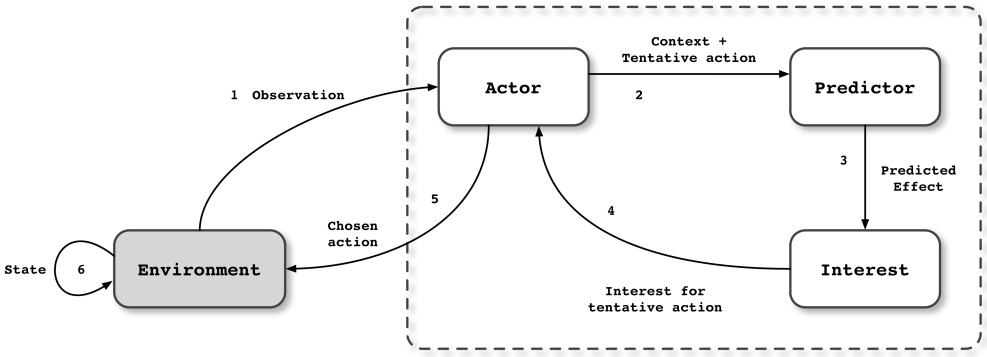


Figure 2: Modular goal-oriented architecture.

Interest Module

Our interest module estimate the interest of each goal the agent conceives.

We define the goal space G as the possible transitions over the space of the state colour boxes. An element of the goal space is called a *task*. In the button experiment, it contains three tasks : no box open $\rightarrow$ no box open, no box open $\rightarrow$ red box open, no box open $\rightarrow$ green box open. If the experiment wasn't divided in trials which reset the robot, we would also have the task : red box open $\rightarrow$ no box open and green box open $\rightarrow$ no box open. The goal space is constructed dynamically as new tasks are observed.

After an action is made, the interest module compares the prediction that was made and the actual output that was observed. The difference, computed as a euclidean distance between vectors describing the box state, is the error in prediction. This signal is monitored across time for each task of the goal space.

Each time a task t is witnessed as being accomplished by the robot, the error relative to the previous prediction is computed, and appended to the list H_t. H_t represents the history of how well the robot was at predicting that the task t was happening.

Given H_t, the interest of task t, i_t, is defined as the first derivative of the last 5 elements of H_t. If H_t contains 0 or one element, i_t is zero.

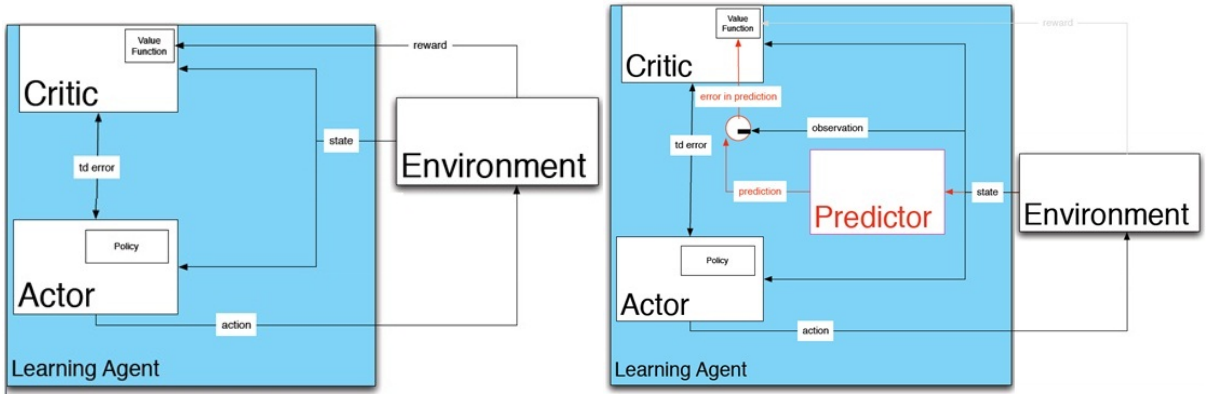
The interest for new tasks (witnessed only one time) is fixed at 50% of the average interest for other tasks. This is helps exploration toward new tasks. The interest for known task t, i_t, is defined as d_t if $d_t > 0$, 0 otherwise, where d_t is the first derivative of the last k elements of H_t. In our implementation, k was fixed at 5. We use a linear least-square fitting on the interest values of the form $\lambda_0 + \lambda_1 \times x$, and define the derivative as equal to λ_1.

3.2 Actor-Critic Architecture

Actor-Critic Architecture with External Reward

According to the standard view of Reinforcement Learning, the agent-environment interaction is described as the interaction between a controller (the agent) and the controlled system (the environment). Figure 3a shows the extrinsically motivated reinforcement learning in which the environment outputs a specialized reward signal, which is taken into account by the critic in the environment that evaluates (usually with a scalar reward value) the agent s behaviour. The agent learns to improve its skill in controlling the environment in the sense of learning how to increase the total amount of reward it receives over time from the critic. The critic updates the valuation of the policy based on the temporal difference error. An ϵ-Greedy algorithm so that the agent explores the environment. Contrarily to other methods, the Temporal-Difference methods are non-model based approaches that found ground from neurosciences Schultz et al. [1997]. The neuromodulator dopamine has long been associated with reward learning.

In our experiment scenario, the environment configurations/states are described in both an state space X (position of the red and green button), and an operational/task space Y (if the red and green spheres have appeared). For given configurations $(x_1, y_1) \in X \times Y$, an action $a \in A$ allows a transition



(a) Extrinsically Motivated Learning

(b) Proposed Prediction Based Intrinsically Motivated Learning

Figure 3: Actor-Critic algorithms for reinforcement learning

towards the new states $(x_2, y_2) \in X \times Y$. The association $(x_1, y_1, a) \mapsto (x_2, y_2)$ corresponds to a learning exemplar that will be memorised, and the goal of our system is to learn both the forward model of the mapping $l : (x_1, y_1, a) \mapsto (x_2, y_2)$.

Actor-Critic Architecture with Intrinsic Reward

An implementation of the Temporal-Difference approach is through the Actor-Critic architecture (fig. 3b). It mostly uses two different entities :

- the actor who decides which action will be taken at every timestep
- the critic who computes the value of each state of the environment with respect to the given task

Both the actor and the critic are updated at every timestep, with respect to a TD error that takes into account the reward received:

$$tdError \leftarrow externalReward + \gamma(critic(x_1, y_1) - critic(x_2, y_2)) \tag{1}$$

where $externalReward$ is given by the environment

In the standard view of Reinforcement Learning, the reward comes from the environment directly, whereas in the intrinsic motivation framework, we will link the reward, not directly to the environment, but to a measure of interest. Our learning agent learns to interact with its environment independently, so as to learn as much as possible. It will therefore try different actions to learn their effects. The goal of the agent is to improve its knowledge, therefore it will focus on the tasks that it does not master yet. The interest of the agent is maximal in situations where it makes mistakes, as these are the situations he will get more information to improve its model. Therefore, we define a predictor as a mapping $(x_1, y_1, a) \mapsto (x_2, y_2)$, which is the model that the robot has of the real mapping l. The predictor is updated at each timestep, according to the observations of the state transitions, as described in fig 4. The reward is then defined with respect to the prediction error:

$$tdError \leftarrow internalReward + \gamma(critic(state) - critic(observation)) \tag{2}$$

where

$$internalReward(x_1, y_1, a, x_2, y_2) = \begin{cases} 1 & \text{if } predictor(x_1, y_1, a) = (x_2, y_2) \\ 0 & \text{otherwise} \end{cases}$$

for every trial
for every timestep
observe from the environment the state s
select an action $selectedA$ with a softmax action
selection
perform action $selectedA$
observe from the environment the new state
$observation$
compute
$tdError \leftarrow internalReward + \gamma(critic(state) - critic(observation))$
where
$internalReward = \begin{cases} 1 \text{ if } predictor(s, selectedA) = observation \\ 0 \text{ otherwise} \end{cases}$
update the critic :
$critic(s) \leftarrow critic(s) + lrC * tdError$
update the actor :
$actor(s, selectedA) \leftarrow actor(s, selectedA) + lrA * tdError$

Figure 4: Pseudo-code of the actor-critic within the framework of Intrinsic Motivation.

3.3 Q-Learning

Q-learning is one among the most often used reinforcement learning algorithms. It is an off-policy temporal difference (TD) control algorithm. This is because the Q-function is updated using a policy that is different from the one used to take actions. One step Q-learning is defined by

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a_t) - Q(s_t, a_t) \right] \quad (3)$$

The learned Q approximates the optimal action-value function independent of the policy being followed.

4 Results

4.1 Button Experiment

With the actor critic algorithm with intrinsic motivation, we tested the competence of the robot in recognising which action to choose for different stimulus. In the testing, we presented the robot with only the red button and plotted in red the prediction error if the red and green spheres would appear. Then we presented the robot with only the green button and plotted the green curve.

Figure 5 shows the average interest level of the agent for each task over multiple trials. Interest for opening the red box is expressed earlier, on average than the green box, and disappears quickly, as the learning is immediate. Opening the green box is a more difficult task, and as such, interest rise slower and stays longer, as the learning takes more time, and progress is sometimes negative. Figure 6 shows the prediction error for different trials of the prediction based actor-critic learning algorithm. The actions which are encountered initially are repeated many times until the prediction error is reduced to a very low value.

4.2 Rod Toppling Experiment

Figure 7 shows the algorithmic progress over time. The graphs show variation in several parameters of the algorithm. Q-table is the control policy developed by Q reinforcement learner based on internal rewards generated from the predictor. Both Q-table and the predictor map state-action pairs to real values. A linear predictor with a learning rate = 0.2 is used. For each state and the action taken by

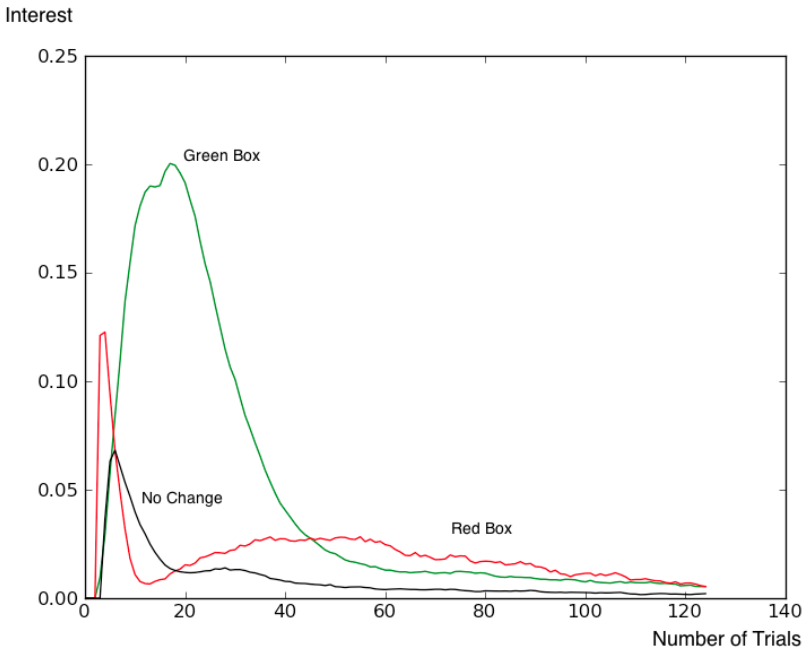


Figure 5: Interest in Goal Oriented Learning Algorithm for different trials

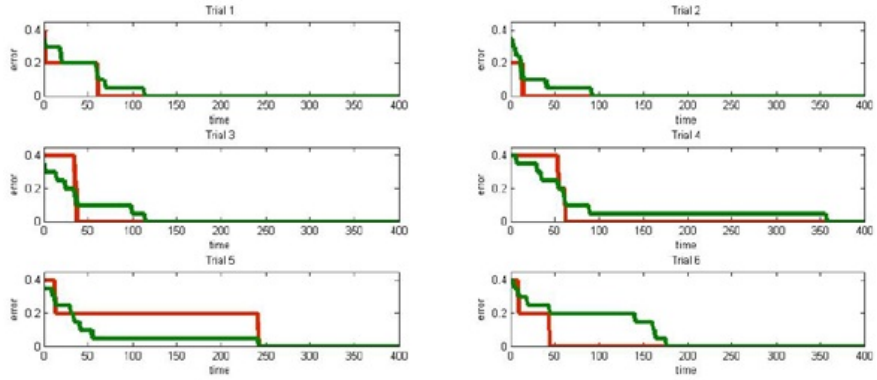


Figure 6: Prediction Error for Prediction Based Actor-Critic Learning Algorithm for different trials

the robot's hand, it tries to predict whether the object is standing still (value = 0) or is-toppled. Until it has learnt to predict well, it makes several prediction errors that are used as intrinsic rewards to the Q-learner. Figure 7(a) shows the experiment after the first episode. The state confidence graph shows the prediction confidence at each state of the robot, while the prediction error graph shows the difference in the prediction made and the actual observation. Each episode ends as soon as the robot topples one of the two rods. Figure 7(b) shows the graphs when the robot starts to predict the red rod well (high state confidence) and makes less prediction errors. Figure 7(c) shows the time when the robot starts interacting with the blue rod. Q-table gives the information about the optimal action sequence to take in order to topple the red rod and the blue rod.

5 Discussion

As demonstrated, both of the proposed methods, the Tabular-Q/Actor-Critic algorithm and the goal-oriented approach resulted in the agent being able to learn both of the required task. In the rod-toppling experiment, the agent also learned the easier association much quicker before being intrinsically

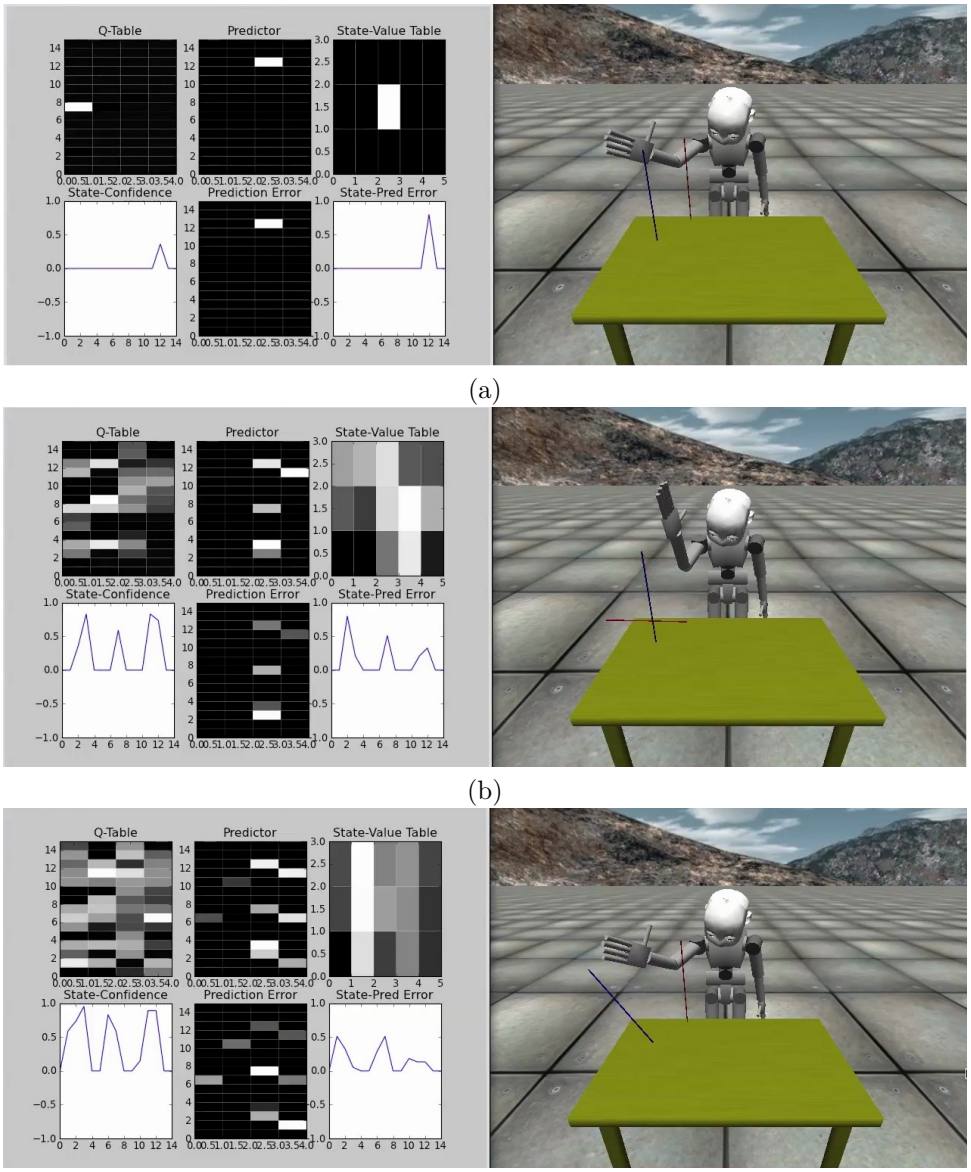


Figure 7: Figures showing the algorithmic progress. The learning is done in an episodic fashion. (a) Robot finds playing with the red rod more interesting. (b) It eventually predicts it and gets bored. (c) It starts to interact with the blue rod.

motivated, via programmed curiosity, to explore the more complex association. An interesting finding with the Actor-Critic algorithm was that the agent's degree of familiarity with the more complex task would temporarily increase above the simple task. This can be explained by noting that the agent would explore the environment sequentially and the location of the objects was assigned randomly. As a result, the agent would first encounter the more complex task before the simpler task. However, eventually the easier task would be mastered faster.

This work could be extended in variations of experiments where the environment is continuous, as well as the actions, but also in the case the world is not deterministic. The framework we chose to explore in this work constrains the exploration to sole intrinsic motivation to explore the world, without prior bias. But this exploration can turn out to be too time-consuming in the case of large or unbounded environment. Coupling intrinsic motivation with bias such as maturational constraints Baranes and

Oudeyer [2010] or social interaction Thomaz [2006] can bootstrap the learning of an open-ended repertoire of skills.

References

Adrien Baranes and Pierre-Yves Oudeyer. Riac: Robust intrinsically motivated active learning. In *Proc. of the IEEE International Conference on Learning and Development.*, 2009.

Adrien Baranes and Pierre-Yves Oudeyer. Maturationaly-constrained competence-based intrinsically motivated learning. In *Proceeding of the IEEE International Conference on Development and Learning (ICDL)*, 2010.

Andrew G. Barto, S Singh, and N. Chenatez. Intrinsically motivated learning of hierarchical collections of skills. In *Proc. 3rd Int. Conf. Development Learn.*, pages 112–119, San Diego, CA, 2004.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

E.L. Deci and Richard M. Ryan. *Intrinsic Motivation and self-determination in human behavior*. Plenum Press, New York, 1985.

V. Fedorov. *Theory of Optimal Experiment*. Academic Press, Inc., New York, NY, 1972.

M. Lopes and Pierre-Yves Oudeyer. Active learning and intrinsically motivated exploration in robots: Advances and challenges (guest editorial). *IEEE Transactions on Autonomous Mental Development*, 2(2):65–69, 2010.

Pierre-Yves Oudeyer and Frederic Kaplan. How can we define intrinsic motivations ? In *Proc. Of the 8th Conf. On Epigenetic Robotics.*, 2008.

Pierre-Yves Oudeyer, Frederic Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):pp. 265–286, 2007.

N. Roy and A. McCallum. Towards optimal active learning through sampling estimation of error reduction. In *Proc. 18th Int. Conf. Mach. Learn.*, volume 1, pages 143–160, 2001.

Richard M. Ryan and Edward L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54 – 67, 2000.

M. Schembri, M. Mirolli, and Gianluca Baldassarre. Evolving internal reinforcers for an intrinsically motivated reinforcement learning robot. In Y. Demeris, B. Scasselati, and D. Mareschal, editors, *Proceedings of the 6the IEEE International Conference on Development and Learning (ICDL07)*, 2007.

J. Schmidhuber. Formal theory of creativity. *IEEE Transation on Autonomous Mental Development*, 2 (3):230–247, 2010.

W. Schultz, P. Dayan, and P. Montague. A neural substrate of prediction and reward. *Science*, 275: 1593–1599, 1997.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: an introduction*. MIT Press, 1998. URL `http://webdocs.cs.ualberta.ca/~sutton/book/the-book.html`.

Andrea L. Thomaz. *Socially Guided Machine Learning*. PhD thesis, MIT, 5 2006. URL `http://www.cc.gatech.edu/~athomaz/pubs.html`.

J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen. Autonomous mental development by robots and animals. *Science*, 291(599-600), 2001.